МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

МЕТОДИЧНІ ВКАЗІВКИ

до лабораторної роботи «Підготовка видання для електронної публікації в репозитарії»

з курсу «Додрукарське опрацювання інформації»
для студентів спеціальності 186 «Видавництво та поліграфія»

Затверджено
редакційно-видавничою
радою університету,
протокол № 1 від 30.01.2018 р.

Харків
НТУ «ХПІ»
2018

Методичні вказівки до лабораторної роботи «Підготовка видання для електронної публікації в репозитарії» з курсу «Додрукарське опрацювання інформації» для студентів спеціальності 186 «Видавництво та поліграфія» / уклад. М. І. Безменов, Л. Б. Кащеєв – Харків : НТУ «ХПІ», 2018. – 16 с.

Укладачі: М. І. Безменов,
Л. Б. Кащеєв

Рецензент В. Л. Лісицький

Кафедра системного аналізу та інформаційно-аналітичних технологій

ВСТУП

Однією з найважливіших тенденцій нашого часу є полегшення доступу до інформації за допомогою глобальної інформаційної мережі Інтернет та електронних носіїв. Очевидно, що джерела інформації (книги, журнали, статті, брошури) в такому випадку повинні бути представлені електронному вигляді (у формі файлів стандартних форматів).

Збереження нової інформації в електронній бібліотеці не викликає труднощів: підготовлене видання не тільки друкується на паперових носіях, а й виставляється в електронному вигляді на відповідному сервері. Складніше йде справа з виданнями минулих років – електронні варіанти їх текстів втрачені або взагалі не існували. Друковані варіанти видань не завжди мають високу якість, часто невдало зшиті і обрізані, на сторінках книг і журналів присутні дефекти. При цьому джерело може мати історичну цінність, через що його не можна розшити або розірвати на сторінки.

Метою лабораторної роботи є відпрацювання питань підготовки однієї або декількох статей журналу або збірника для електронної публікації в форматі pdf.

1. ПОСТАНОВКА ЗАДАЧІ

Підготовка публікації для зберігання в електронному репозитарії повинна пройти наступні етапи:

1. Сканування сторінок або розворотів зі збереженням їх у растровому файлі.
2. Графічна обробка зображень (очищення від дефектів, підвищення контрастності, якщо треба – поворот зображення для вирівнювання рядків по горизонталі).
3. Розпізнавання та корекція тексту.
4. Формування єдиного документа зі сканованих зображень, наприклад, doc-файлу в редакторі Microsoft Word.
5. Експорт документа в формат pdf.

6. Складання текстового файлу з анотаціями статей в pdf документі.

Розглянемо порядок дій по кожному з пунктів цього плану.

2. ТЕОРЕТИЧНІ ОСНОВИ
2.1 ФОРМАТИ ЕЛЕКТРОННИХ ВИДАНЬ

Зараз основними стандартами для електронних публікацій є формати pdf, djvu, fb2 і epub. Дані про ці формати наведені в табл. 2.1.

Таблиця 2.1 – Основні формати електронних видань

Формат	Рік розробки	Розробник	Основні програми для читання
pdf	1993	Adobe Systems	Adobe Reader, Mozilla Firefox, Foxit Reader, Google Chrome
djvu	1998	AT&T Labs Research	WinDjView, STDU Viewer, DjVuReader
fb2	2004	Д. Грибов	Ice Book Reader, CoolReader, FDREader, AlReader
epub	2007	IDPF	Calibre, FBReader, AIReader

Якщо порівнювати ці формати зі зберіганням книг у форматі txt, який широко використовується в електронних книгах Amazon Kindle, то характерною особливістю форматів pdf, djvu, fb2 і epub є наявність в них інформації як про основний текст, так і про вкладені двійкові файли (рисунки, елементи управління). Очевидно, що для читача присутність текстового контенту виглядає більш переважно: в подібних електронних виданнях можна здійснювати пошук за словами і словосполученнями, фрагменти тексту легко копіювати для цитування. Однак в старих виданнях, коли фрагменти текстового контенту вписані руками (наприклад, формули , слова і символи на

інших мовах), весь текст часто вимушено представляється в бінарному вигляді (у формі растрового зображення). Істотним недоліком такого подання є значно більший обсяг дискового простору для зберігання подібних файлів. Тому бажаним все ж є розпізнавання тексту з подальшою вставкою нерозпізнаних його частин у вигляді об'єктів-малюнків.

2.2 СКАНУВАННЯ СТОРІНОК

Переклад друкованого тексту і зображень в електронну форму називається скануванням. Цей процес здійснюється спеціальним пристроєм для введення інформації – сканером.

Найважливішим елементом будь-якого сканера є світлочутлива матриця. Вона трансформує зміни кольору і яскравості світлового потоку, відбитого від паперу, в аналогові електричні сигнали. Більшість сучасних сканерів для дому та офісу базуються на матрицях двох типів – CCD (Charge Coupled Device) і CIS (Contact Image Sensor). Зовнішня різниця цих пристроїв помітна навіть неозброєним поглядом – CIS-сканер має низький плоский корпус висотою 30-40 мм, а висота аналогічного CCD-сканера не менш 50 мм. Ця різниця зумовлена особливістю освітлення листа-джерела: в CIS-сканері папір висвітлюється лінійкою світлодіодів, в CCD використовується джерело світла, промінь фокусується на кожній точці рядка за допомогою системи дзеркал. Вплив елементів конструкції на результати сканування, переваги і недоліки сканерів наведені в табл. 2.2.

У плані поставленого завдання особливе значення має глибина різкості – наскільки може відстояти від матриці поверхню книги у місця зшивання (палітурки). Притиснути щільно сторінки книги в районі палітурки без руйнування прошивки фізично неможливо. Сканери по-різному зчитують інформацію в цих місцях. Відмінності в якості сканування CCD і CIS сканерів показані на рис. 2.1.

Таблиця 2.2 – Переваги і недоліки CCD- и CIS-сканеров

Тип сканеру	Переваги	Недоліки
CCD	Висока роздільна здатність (недорогі CCD-моделі мають роздільну здатність до 2400 dpi)	Висока вартість (в порівняно з CIS)
	Довгий термін служби ламп	Тривалий прогрів лампи до початку роботи
	Висока якість сканування	Необхідність в додатковому джерелі живлення
	Велика глибина різкості	
CIS	Невеликі габарити	Обмежена роздільна здатність (зазвичай до 1200 dpi)
	Швидкий старт	Невелика глибина різкості
	Низький рівень споживання енергії (в тому числі харчування через USB)	Чутливість до бічного засвічування
	Невисока вартість	

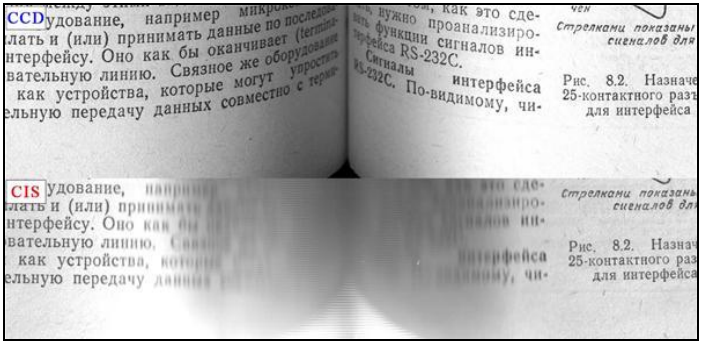

Рисунок 2.1 – Зверху CCD сканер, знизу – CIS сканер

З зображень на рис. 2.1 стає очевидним, що після сканування буде потрібна додаткова обробка зображень. Однак для CIS-сканера вона буде значно складніше.

2.3 РОЗДІЛОВА ЗДАТНІСТЬ ПРИ СКАНУВАННІ

Розділова здатність – це величина, яка визначає кількість елементів растрового зображення (пікселів) на одиницю довжини (або площі). Змінити дозвіл можна при скануванні або при редагуванні зображення в графічному редакторі. Для позначення роздільної здатності сканерів використовують одиницю вимірювання dpi – кількість точок на дюйм (dots per inch).

Як правило, сучасні сканери можуть сканувати зображення в декількох розділових здатностях. Частіше за інших зустрічаються значення 72 dpi (в WEB або презентаціях), 300 dpi (для друку в поліграфії кольорових і тонових ілюстрацій) і 600 dpi (друк креслень).

При підготовці електронного видання слід пам'ятати, що якщо в подальшому планується розпізнавати сканований текст (наприклад, програмою ABBYY FineReader), то висока якість розпізнавання вимагає подання вихідного зображення в розділовий здатності не менше 300 dpi.

2.4 СКАНУВАННЯ В ПРОГРАМІ PHOTOSHOP

Для роботи зі сканером можна використовувати стандартну програму сканування, яка поставляється кожною фірмою-виробником разом зі своїм пристроєм, а можна скористатися універсальними програмами, які працюють майже з усіма типами сканерів. Далі описано, як можна здійснювати сканування за допомогою зовнішніх модулів програми Photoshop.

Сканування полягає у виконанні наступних етапів.

1. Помістити скановану сторінку або розворот лицьовою стороною на скло сканера і вирівняти документ по лінейкам. Бажано не присувати лист до країв скла, оскільки 3-5 мм по краях – «мертва зона» для сканування.

2. У програмі Photoshop вибрати модуль сканування або (в залежності від версії Photoshop) виконати команду File → Import → Twain_32 (Файл → Импортировать → Twain_32) або File → Import →

WIA Support ... (Файл → Импортировать → Поддержка WIA ...). Пункти меню та повідомлення на скріншотах наведені російською мовою – програма Adobe PhotoShop портована на російської мові.

Якщо пристрій сканування не підключений до комп'ютера, подальший діалог настройки обривається, рис. 2.2.

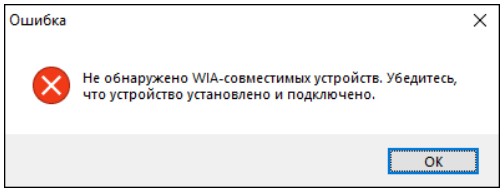

Рисунок 2.2 – Вікно-попередження, що повідомляє про відсутність сканерів, приєднаних до комп'ютера

Якщо з підключенням сканера все в порядку, на екран виводиться діалогове вікно, рис. 2.3.

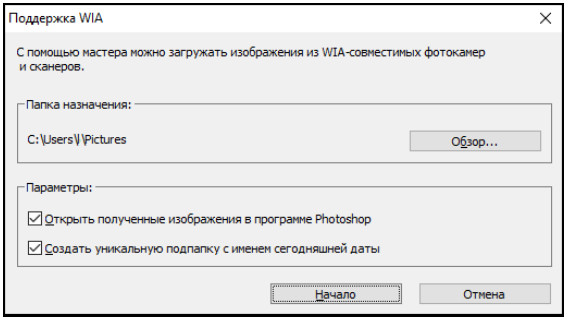

Рисунок 2.3 – Вікно для налаштування шляхів збереження відсканованих зображень

3. Клацнути по кнопці «Початок» – програма виводить список сканерів, які доступні для підключення в даний момент.

4. Якщо мова йде про сканування сторінки в цілому, то в наступному діалоговому вікні, рис. 2.4, натиснути кнопку «Сканувати», коли з усього документа передбачається сканувати лише фрагмент – натиснути кнопку «Перегляд» і вибрати область сканування.

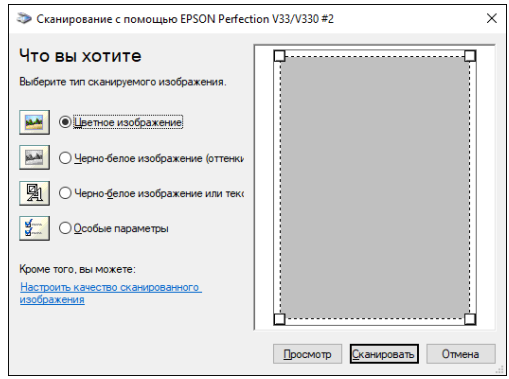

Рисунок 2.4 – Вікно для початку сканування документа

При скануванні першого розвороту слід натиснути «Налаштувати якість зображення, що сканує» і вибрати необхідний дозвіл, рис. 2.5. Для подальших листів або розворотів установки зберігаються.

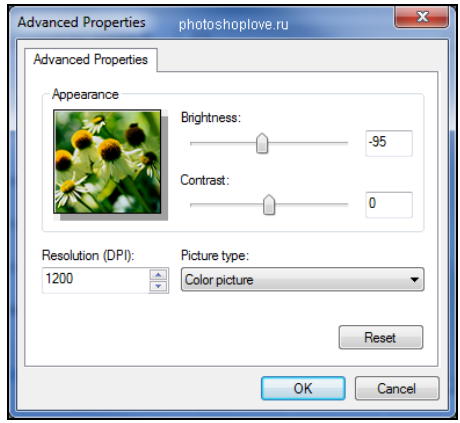

Рисунок 2.5 – Вікно настройки якості сканування

Сканування високоякісних текстових документів можна вести в режимі «Чорно-білого документа або тексту», сканування не надто якісних сторінок краще здійснювати в режимі «Чорно-біле зображення (відтінки сірого)». Якщо папір журналу або книги пожовклий від часу, його треба сканувати як «Кольорове зображення».

2.5 ОБРОБКА ЗОБРАЖЕНЬ В PHOTOSHOP

Зміна контрастності, знебарвлення і збереження отриманих зображень сторінок в редакторі Photoshop можна виконати в напівавтоматичному режимі, але для цього потрібно привести кожну сторінку видання до єдиного вигляду. З цією метою пропонується виконати наступну послідовність дій.

1. Створити «еталонну сторінку». Для цього береться одна з цих сторінок і очищається від тексту. По краях сторінки за допомогою інструменту «Штамп» редактора Photoshop стираються плями, затемнення (особливо на переломі палітурки сторінок) та інші дефекти. Виставляється бажана розділова здатність і розмір документу.

2. Відкрити всі сторінки. По черзі на кожній сторінці виділяється середня текстова частина, яка копіюється і вставляється поверх «еталонної сторінки». Тим самим досягається відмова від найбільш забруднених крайових ділянок листа.

3. Покласти прошитий журнал або книгу строго уздовж напрямних лінійок сканера вкрай складно. Тому, швидше за все, кожну сторінку доведеться трохи повернути. Для цього потрібно в основному меню редактора Photoshop при виділеному шарі виконати команду Edit → Free Transform (Редактирование → Трансформирование). Виставляти рядки строго горизонтально немає необхідності – програми для розпізнавання тексту відпрацьовують невеликий перекіс рядків.

4. У разі помітних дефектів, які при розпізнаванні будуть сприйматися як букви або символи, їх також бажано видалити, наприклад, за допомогою інструментів «Штамп» або «Ластик».

5. Об'єднати шари відредагованого зображення і зберегти його в jpg файлі (кожну сторінку в окремому файлі).

Схематично цей процес редагування показаний на рис. 2.6.

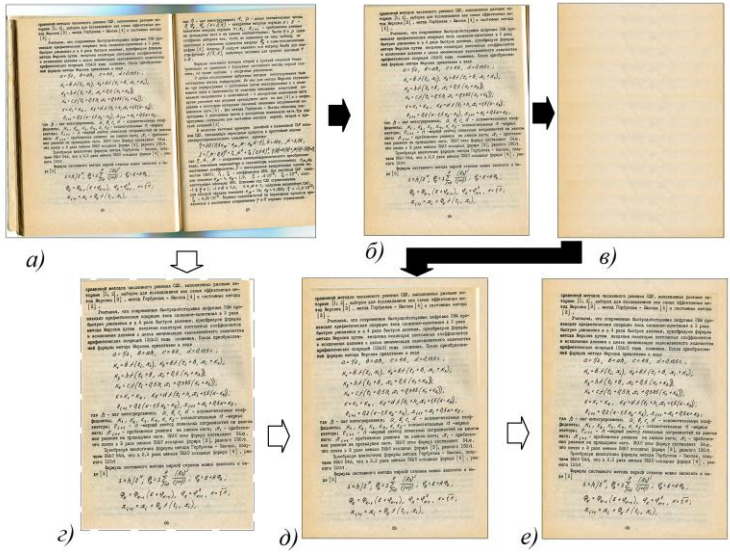

Рисунок 2.6 – Послідовність обробки відсканованого зображення

На рис. 2.6 відображена схема редагування відсканованих сторінок: *а)* відсканований розворот; *б)* одна сторінка розвороту; *в)* на одній сторінці стертий вихідний текст, тепер це «еталонний фон»; *г)* з вихідного розвороту скопійований фрагмент з текстом; *д)* фрагмент накладено на «еталонний фон», *е)* фрагмент повернуть і зачищений по краях. На рисунку послідовність *а–б–в* – створення фонової основи, *а–г–д–е* – формування виправленого зображення на очищеному тлі

Далі необхідно виконати рутинну, багато разів повторювану роботу в редакторі Photoshop.

Після попереднього редагування всіх розворотів або сторінок видання всі отримані сторінки слід знебарвити (перетворити в градації сірого) і збільшити їх контрастність для підвищення якості розпізнавання в Abbyy FineRider.

З цією метою напишемо наступний макрос.

1. Завантажуємо одну з оброблюваних сторінок.

2. Переходимо у вікно «Операции» (Actions) (якщо воно не відкрито, то попередньо відкриємо його: Окно → Операции (Window → Actions). На нижній рамці вікна вибираємо кнопку «Создает новую операцию» (Create new action). Ця кнопка має вигляд листочка з загнутим куточком. Результатом є відкриття вікна «оформлення» макросу (рис. 2.7).

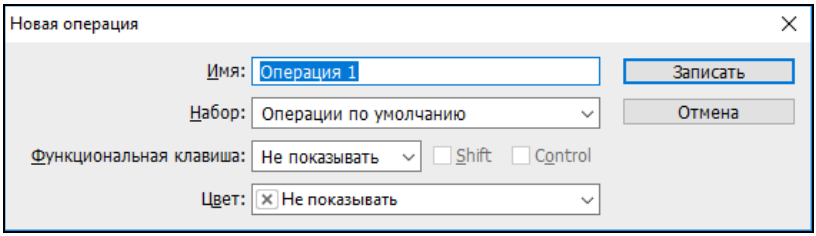

Рисунок 2.7 – Вікно для присвоєння макросу імені та гарячої клавіші

3. У вікні необхідно вибрати гарячу клавішу для запуску макросу. Наприклад, вибираємо у спливаючому списку (де зараз рядок «Не показувати») функціональну клавішу F2 і ставимо галочку в пункті Shift; рис. 2.8.

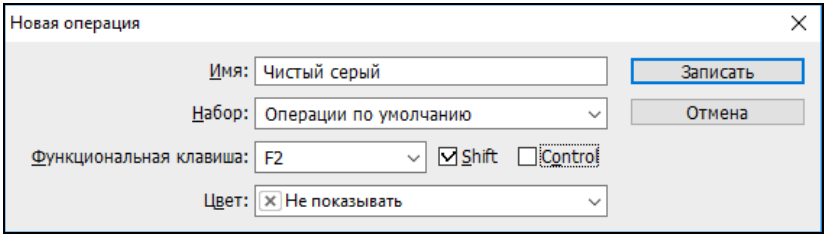

Рисунок 2.8 – Макросу присвоєно ім'я та гарячі клавіші

Ім'я можна і не привласнювати, але при наявності великої кількості схожих макросів в списках «История» (History) і «Операции» (Actions) краще мати осмислені назви, щоб простежити, яка операція тільки що виконувалася.

4. Після натискання кнопки «Записать» починається запис макросу, про що свідчить те, що значок у вигляді сірого кола в нижній смузі вікна «Операции» стає червоним. Поспішати при запису операцій немає сенсу: час не фіксується, записується тільки послідовність дій. Тому краще виконати всі операції гранично чітко, щоб не довелося перезаписувати макрос. При цьому пропонується виконати такі дії.

4.1. Збільшуємо контрастність зображення до межі (до 100), рис. 2.9.

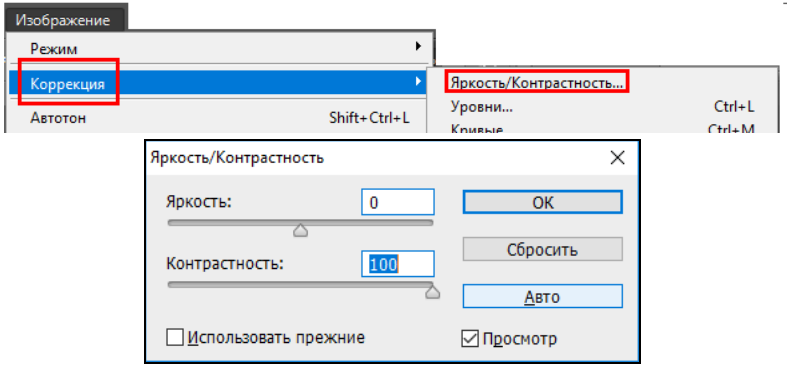

Рисунок 2.9 – Збільшення контрастності зображення

4.2. Переводимо кольорове зображення в градації сірого кольору (рис. 2.10), після чого підтверджуємо втрату квітів натисканням кнопки «Отменить».

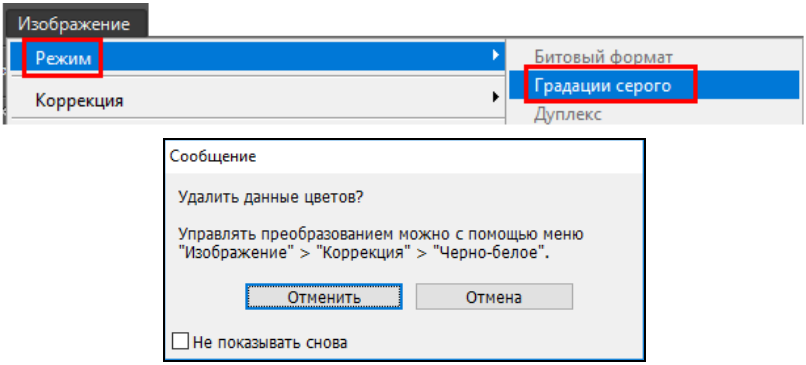

Рисунок 2.10 – Переклад зображення в градації сірого кольору

4.3-4.5. Ще кілька разів (зазвичай два-три) збільшуємо контрастність до межі, по суті – повторюємо пункт 4.1.

5. Завершуємо запис, що забезпечується натисканням квадратної кнопки, розташованої лівіше індикатора записи на нижній рамці вікна «Операції». Макрос створений і готовий до застосування.

6. Зберігаємо відредаговану сторінку і завантажуємо нову – ту, яку необхідно перетворити в градації сірого кольору і контрастність якої повинна бути збільшена майже до чорно-білого виду. При натисканні Shift + F2 – вся послідовність операцій виконується автоматично.

Цілком реальна ситуація, коли робота виявиться занадто об'ємною, і її не вдасться виконати за один сеанс роботи в Photoshop. Більш того, при тимчасове припинення роботи на комп'ютері, можливо, в Photoshop буде працювати інший користувач зі своїм набором акцій (макросів). Для вирішення цієї проблеми створюємо новий набір макросів (set) і перетягуємо мишею наш макрос в папку цього набору.

3. РОЗПІЗНАВАННЯ ТЕКСТУ

Для розпізнавання тексту зазвичай користуються програмою Abbyy FineReader. При запуску ця програма запитує, звідки буде взято вихідний документ. У розглянутому вище випадку були підготовлені jpg файли, так що вибираємо «З файлу (PDF / зображення) в документ Microsoft Word».

Після відкриття файлу програма автоматично починає процес розпізнавання. Якщо в результаті отримано документ з великою кількістю помилок (наприклад, рис. 3.2), потрібно налаштувати програму FineReader на поточне завдання. Зокрема, можна скасувати режим «Автоматичний вибір» мови і задати конкретні мови, на яких написаний розпізнається текст.

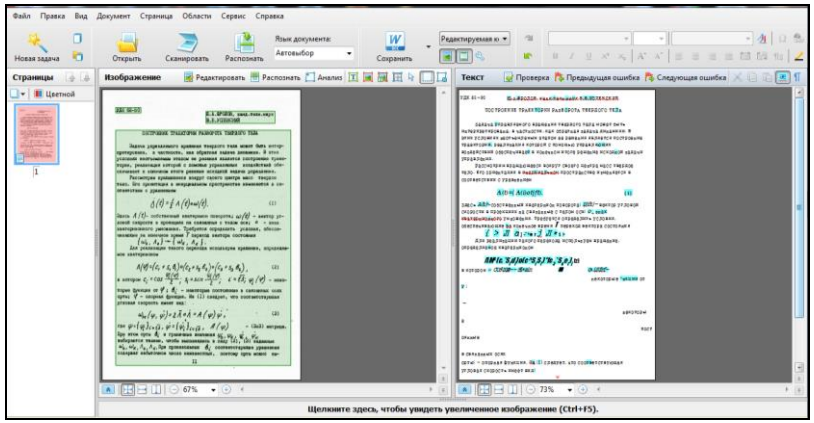

Рисунок 3.2 – Текст розпізнано, але число помилок дуже велика

4. ОТРИМАННЯ PDF-ФАЙЛА

Після закінчення розпізнавання тексту слід зібрати всі сторінки в єдиний документ і здійснити текстове редагування. Навіть при хорошій якості розпізнавання в тексті залишаються граматичні помилки, на стиках сторінок часто є неприбрані переноси, присутні помилки в розпізнаванні формул і так далі. Після отримання остаточного тексту слід в редакторі Word виконати команду «Файл → Зберегти як» і вибрати для збереження pdf формат. Можна отримати pdf-файл, якщо при відсутності принтера направити документ на друк, в діалоговому вікні буде запропоновано вибрати шлях збереження pdf-файлу.

5. СКЛАДАННЯ ФАЙЛА ЗМІСТУ

Зберігання в репозитарії збірників статей має на увазі не тільки запис електронних версій видань, а й складання файлу-вмісту для звернення до них. (В різних репозитаріях вимоги до формату супроводжуючих документів можуть суттєво відрізнятися).

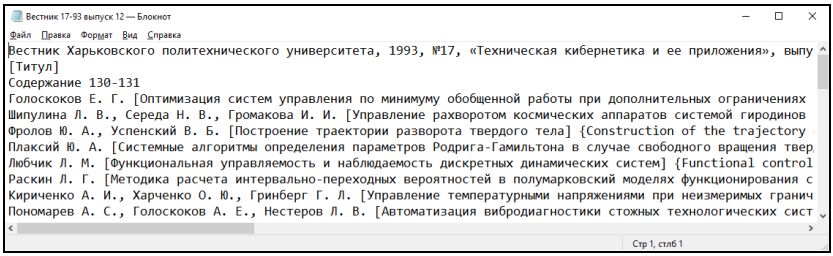

Рисунок 5.1 – Оформлення текстового файлу-вмісту

Файл-вміст оформляється у вигляді тексту. Формат файлу (точніше, приклад формату) наведено на рис. 5.1:

автори [назва мовою оригіналу], {назва англійською мовою}, сторінки

6. ЗАВДАННЯ НА ЛАБОРАТОРНУ РОБОТУ

В ході роботи потрібно відсканувати, розпізнати і підготувати у вигляді pdf-файлу фрагмент з 10 сторінок з Вісника «Технічна кібернетика та її застосування», випуск 12 (17’93):

Варіант 1 – сторінки 40-49, варіант 2 – сторінки 50-51, варіант 3 – сторінки 60-61, варіант 4 – сторінки 70-71, варіант 5 – сторінки 80-81, варіант 6 – сторінки 90-91 і так далі.

Розроблений pdf-файл супроводити текстовим фрагментом змісту.

СПИСОК ЛІТЕРАТУРИ

1. Пічугін М. Ф., Канкін І. О., Воротніков В. В. Комп’ютерна графіка : навч. посіб. Київ : Центр учбової літератури, 2013. 346 с.

3. Кащеев Л. Б., Безменов Н. И., Чех Л. И. Анализ возможности использования графических форматов для экономичного хранения изображений видеоряда. *Вестник Харьковского политехнического института. Серия: Техническая кибернетика и ее приложения, вып. 12*. 1993, № 17. с. 90–94.